(19) Europäisch s Patentamt
European Patent Offic
Office urop é n des brev ts

(11) Publication number : 0 674 263 A1

(12) EUROPEAN PATENT APPLICATION

(21) Application number : 95301345.5

(51) Int. Cl.[6] : G06F 11/20, G06F 11/14

(22) Date of filing : 02.03.95

(30) Priority : 21.03.94 US 215238

(43) Date of publication of application : 27.09.95 Bulletin 95/39

(84) Designated Contracting States : DE FR GB

(71) Applicant : International Business Machines Corporation
Old Orchard Road
Armonk, N.Y. 10504 (US)

(72) Inventor : Micka, William Frank
3921 E La Espalda
Tucson, Arizona 85718 (US)
Inventor : Shomler, Robert Wesley
17015 Piedmont Ct.
Morgan Hill, California 95037 (US)

(74) Representative : Davies, Simon Robert
I B M
UK Intellectual Property Department
Hursley Park
Winchester, Hampshire SO21 2JN (GB)

(54) Asynchronous remote data copying.

(57) Asynchronous remote data duplexing at a distant location is performed from copies based at a primary site storage subsystem using first and second pluralities of subsystems 12, 14 at primary and remote sites respectively. Each of the first plurality of subsystems is independently coupled to one or more of the second plurality of subsystems. The first plurality of subsystems are interconnected and the second plurality of subsystems are also interconnected. The method utilizes checkpoint messages to maintain sequence integrity between the first and second plurality of subsystems without the use of a centralized communications service.

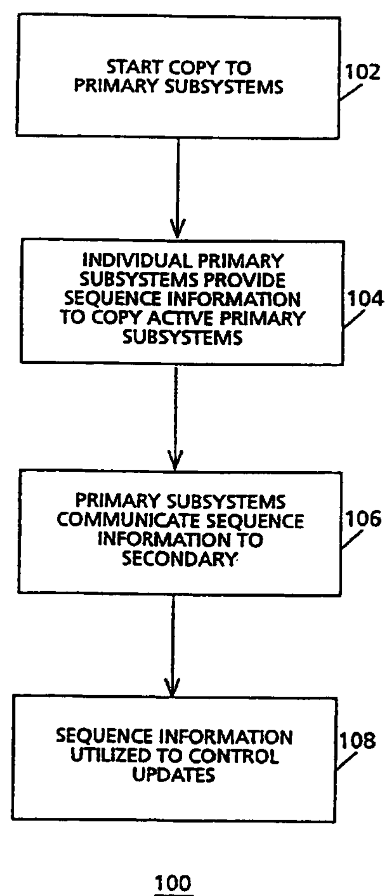


Figure 3

EP 0 674 263 A1

The present invention relates to asynchronous r mote data copying or dupl xing to provid data pres rvation in an inf rmation handling system, wh reby information from a primary site storage subsyst m can be copied to a remote l cation.

Data copying is one form of data preservation in an information handling or computer system. However, data preservation via data copying must take many factors into account. This is of special significance where it is anticipated that data copied and stored at a remote site would be the repository for any continued interaction with the data should the work and data of a primary site become unavailable. The factors of interest in copying include the protection domain (system and/or environmental failure or device and/or media failure), data loss (no loss/partial loss), time where copying occurs as related to the occurrence of other data and processes (point in time/real time), the degree of disruption to applications executing on said computer, and whether the copy is application or storage system based. With regard to the last factor, application based copying involves log files, data files, and program routines while storage based copying involves an understanding of direct access storage device (DASD) addresses with no knowledge of data types or application use of the data.

Real-time remote data duplexing systems require some means to ensure update sequence integrity as write updates to the secondary or remote DASD data copy. One way to accomplish this is to provide a synchronous system to control the DASD subsystems. In such a system, the primary DASD write operation does not complete until a copy of that data has been confirmed at a secondary location. The problem with such synchronous systems is that they slow down the overall operation of the duplexing system.

Asynchronous copy systems accomplish sequence integrity through a centralization and consolidation of data communications between primary and secondary DASD subsystems through a central communications system. In such systems, a system at the primary site can determine the sequence among different update write operations among all DASD subsystems at the primary site and communicate that information to the DASD subsystem at the remote site. The secondary subsystem in turn uses the sequence information from the primary to control the application of update data to the secondary DASD data copy. Known asynchronous copy systems that utilize centralized data communications are described below.

McIlvain and Shomler, U.S. patent application number 08/036,017 entitled "Method and Means for Multi-System Remote Data Duplexing and Recovery" (EPA 617362) describes the us of a store and forward message interface at the DASD storag management level betwe n a source of update copies and a remote site in a host to host coupling in which the differ nce in update completeness or loss of the sequence of write updates could be complet ly specified in the v nt of interruption.

Cheffetz, et al., U.S. Patent 5,133,065 entitled "Backup Computer Program for Networks" issu d July 21, 1992, discloses a local area network (LAN) having a file server to which each local node creates and transmits a list of local files to be backed-up. Such remote generation reduces the traffic where a network server initiates the list creation and file copying activity. Arguably, art published before this reference taught centrally administered file selection. This resulted in compromises to local node security and overuse of the server. This is presumptively avoided by Cheffetz's local node generated lists and remission of the lists to the file server.

Beale, et al., U.S. Patent 5,155,845 entitled "Data Storage System for Providing Redundant Copies of Data on Different Disk Drives", copies variable length records (CKD) on two or more external stores by causing a write to be processed by the first storage controller and be communicated in parallel over a direct link (broad band path) to the second storage controller obviating the path length limitation between the primary and remote copy sites. Such a limitation is occasioned by the fact that CKD demand/response architecture is length limited to in the range of 150 meters.

Another example of an asynchronous system that utilizes a centralized system is disclosed in U.S patent application entitled "Remote Data Duplexing Asynchronous Information Packet Message, by Micka et. al. (EPA 602822). This system discloses a system for asynchronously duplexing direct access storage device (DASD) data in a plurality of DASD subsystems and has the advantage of decoupling the data duplexing operation from the DASD write I/O operation. This ensures the write does not incur unnecessary wait states in the subsystem. By establishing a sequence checkpoint at which time a set of information packets are grouped together and processed as a single sequence unit, this decoupling and independent operation takes place. Through this independence, data copying to a secondary location can take place without affecting the performance of the subsystems and also not affecting the corresponding integrity of the data that is being updated.

There are systems in use in which there is no centralized communication service between the primary and secondary locations. Oftentimes in such system configurations each primary subsystem has a direct independent link to a selected secondary subsystem. In such a system that includes subsystems providing sequence consistent asynchronous write operations can not be address d utilizing known asynchronous copy sch mes.

Accordingly, the invention provid s a method of operating an asynchronous remote data copy system

including a primary site having a first plurality of subsystems interconnected by a first coupling means, and a secondary site remot from the primary site having a second plurality of subsystems interconnected by a second coupling means, each of the second plurality of subsystems being independently coupled to a counterpart one of the first plurality of subsystems, said method comprising the steps of:

in the first plurality of subsystems;

sending a checkpoint signal to each of the first plurality of subsystems;

sending updated data and the checkpoint signal to each of the counterpart coupled second plurality of subsystems; and

in the second plurality of subsystems;

receiving the updated data and checkpoint signals; and

coordinating the writing of the updated data based upon the checkpoint signals.

Such method provides independent links between primary and secondary subsystems, with no central communications system, and provides for sequence consistent remote copying from one set of multiple subsystems at one location to a second set of multiple subsystems at a second subsystem. The asynchronous copy system is simple, cost effective and does not significantly impede the overall operation of the subsystems.

In a preferred embodiment the method further comprises the step of distributing a sequence signal at the primary site, said checkpoint signal having a predetermined relationship with the sequence signal. Such a sequence signal is used to ensure synchronisation between the various subsystems. However, if the network can ensure quick and reliable distribution of the checkpoint messages, then in some situations it is possible to dispense with the sequence signal.

In the preferred embodiment, the method further comprises the steps of: activating each of the first plurality of subsystems to communicate with the other subsystems in said first plurality of subsystems; activating each of the first plurality of subsystems to communicate with its counterpart coupled subsystem in the second plurality of subsystems; building at least one configuration table in each of the first plurality of subsystems such that each of the first plurality of subsystems can identify all of the other subsystems in said first plurality of subsystems; and synchronizing the first plurality of subsystems; the coordinating step further comprises the steps of: receiving copy active messages in the second plurality of subsystems; building copy active tables in the second plurality of subsystems; synchronizing copy operations of the second plurality of subsystems; receiving the ch ckpoint messages in the second plurality of subsystems; perf rming a rendezvous f r all checkpoint m ssages in the second plurality of subsystems; and applying the updated data at the second plurality of subsystems.

Th m thod of th preferred embodiment furth r comprises the steps of: causing each of said first plurality of subsystems to asynchronously g nerate a sequence of updates; ordering each of said sequ nces in accordance with the received sequence signal and checkpoint signal; asynchronously communicating sequences of updates from each of the first plurality of subsystems into a buffered portion of a counterpart subsystem; and applying the buffered sequences at the second plurality of subsystems as a function of each checkpoint message.

The invention additionally provides a method of operating an asynchronous remote data copy system including a primary site having a first plurality of subsystems interconnected by a first coupling means, and a secondary site remote from the primary site having a second plurality of subsystems interconnected by a second coupling means, each of the second plurality of subsystems being independently coupled to a counterpart one of the first plurality of subsystems, said method comprising the steps of:

(a) at the primary site responsive to initiation of a start copy operation, ascertaining at each subsystem the subset of the plurality of DASD subsystems forming the copyset group, designating one of the plurality of subsystems as a clocking and checkpoint message source, each checkpoint message including a sequence clock value and an increased sequence number;

(b) at the secondary site, repeating step (a) for counterpart ones of the DASD subsystems;

(c) at the primary site, periodically generating clocking signals and checkpoint messages by said designated subsystem and broadcasting said signals and messages as they occur to other subsystems forming the copyset group including itself, at each subsystem in the copyset group, and

(1) asynchronously forming a local sequence of updated records,

(2) embedding said signals and messages into the sequence to form a time discriminated total ordering of updated records, and

(3) remitting at least a portion of the sequence to a buffer portion of the counterpart DASD subsystem at the secondary site; and

(d) at the secondary site, applying a checkpoint message to the designated subsystem operative as a synchronizing source by each subsystem in the copyset group and writing the sequences or portions thereof stored in the buffered portions of said subsyst ms to DASD only upon a signal from said designated subsystem indicative of its receipt of all send messages.

Each checkpoint m ssage signifi s that all DASD records with a clock signal with a sequence number that is l ss than the checkpoint message se-

quence clock value hav been transmitt d to the counterpart secondary subsyst m.

The invention further provides an asynchronous remot data copy system including a primary site having a first plurality of subsystems interconnected by a first coupling means, and a secondary site remote from the primary site having a second plurality of subsystems interconnected by a second coupling means, each of the second plurality of subsystems being independently coupled to a counterpart one of the first plurality of subsystems,

the first plurality of subsystems including means for sending a checkpoint signal to each of the first plurality of subsystems; and means for sending updated data and the checkpoint signal to each of the counterpart coupled second plurality of subsystems;

and the second plurality of subsystems including means for receiving the updated data and checkpoint signals; and means for coordinating the writing of the updated data based upon the checkpoint signals.

Preferably, the checkpoint signal sending means comprises: means for sending a checkpoint message to the first plurality of subsystems; and means responsive to the checkpoint message sending means for inserting the checkpoint message into an update data sequence from each of said first plurality of subsystems to its counterpart subsystem.

Thus the above method and means may be used for the remote copying of data at a secondary DASD subsystems. Typically in such an implementation, the secondary DASD subsystems are peer coupled to counterpart ones of a primary host attached DASD subsystems. A start copy operation is initiated at the host by a message broadcast to all primary DASD subsystems. Each primary builds a configuration table using a local area network or other suitable means as a set associative device for table building. This table identifies all the primary subsystem participants. Also, each primary synchronizes its local clock or other time reference to that of a designated primary subsystem, the designated primary being operative as a clocking and checkpoint message source. The designated primary periodically sends checkpoint messages having a sequence time value and an increasing checkpoint sequence number, to each primary subsystem. These are logically inserted into its local copy write record sequence. At the secondary host attached DASD subsystems, one of the secondary subsystems is designated as a local synch source. Each secondary subsystem builds a configuration table of copy active subsystems, and couples to the counterpart primary subsystem. Next, each secondary subsystem asynchronously receives and locally buff rs a copy sequence from its primary count rpart. Each rec iv d s qu nce includes ch ckpoint messages emb dd d therein at th primary after all tim stamped write updated records hav been s nt. R sponsive to receipt of a checkpoint messag , the sec ndary subsystem will remit it to the local synch source. A rendezvous is executed by this source over all the checkpoint messages. Thus, ach secondary subsystem writes from buffer to DASD upon receipt of the rendezvous completion from the synch source. If one or more of the primary DASD subsystems become unavailable, then their counterparts at the secondary site resign from the copy-group. Such occurs only after completion of any updates in progress.

This approach allows for distributed non-central-system operated control of a sequence-consistent real-time asynchronous data copy from a set of DASD subsystems at a primary location to a set of DASD subsystems at a secondary location, with primary subsystems separately and independently connected to secondary subsystems. This enables update sequence integrity of a data copy at a plurality of subsystems remote from a source of asynchronously independently generated sequence of write operations, there being a first plurality of subsystems at a primary site, the first plurality of subsystems being interconnected by a first coupling means, and a second plurality of subsystems at a site remote from the primary site, the second plurality of subsystems being interconnected by a second coupling means, each of the second plurality of subsystems being independently coupled to one of the first plurality of subsystems.

Thus in a system for remote copying of data at a secondary site having a coupling plurality of storage device (DAS) subsystems that are interconnected by a first couplingmeans, said secondary DASD subsystems being coupled to counterpart ones of a plurality of DASD subsystems at a remote primary site, said secondary DASD subsystems being interconnected by a second coupling means, said system including means at the primary site for initiating a start copy operation, a method is provided comprising the steps of:

(a) at the primary and secondary sites, forming m copyset groups of DASD subsystems respectively;

(b) at the primary site:

(1) causing each of the m subsystems to asynchronously generate a sequence of updated write records;

(2) ordering each of said sequences by embedding common clock values and a periodic checkpoint message with a common clock value and increasing a sequence number in each of said sequences,

(3) coupling count rpart DASD subsystems between the sites and asynchronously communicating s quences from ach of the primary sit subsystems into a buffered portion of a counterpart s condary site subsystem; and

(c) at the s condary site, writing the buffered sequences to DASD in the counterpart subsystems as a function of each checkpoint messag .

An embodiment of th invention will now be described in detail by way of example only, with reference to the following drawings:

Figure 1 is a conventional remote data copy system configuration;

Figure 2 is remote dual copy system configured in accordance with the present invention;

Figure 3 is a flow chart showing the general operation of the remote dual copying of Figure 2; and

Figure 3A-3D are more detailed flow charts showing the operation of the remote dual copying system of Figure 2.

To perform asynchronous remote copying, (1) the sequence of data updates must be determinable at a local site; (2) that sequence must be communicable by the local site to a remote site; and (3) the remote site must be able to use the sequence to control the updating at the remote site.

As has been before mentioned, prior art asynchronous copy systems which include multiple subsystem require a central system for providing the appropriate sequences of data. However, system configurations may be found in which it is not desired to pass update data between the primary and secondary locations through a centralized communication service. Rather in those configurations it is desired to directly connect DASD subsystems at the primary and secondary sites via independent subsystem-to-subsystem communication links. Such a system 10 is shown in Figure 1. The system 10 includes a host 11 which provides data to primary subsystems 12'. As is shown each of the links between a primary DASD subsystem 12' and its peer coupled secondary DASD subsystem 14' is independent. As a result, this type of system would inherently be incapable of write operations that are sequentially consistent.

To more specifically describe the problem, consider as an example a sequence of three writes from a conventional database management system (DBMS) as it is about to commit a transaction. The example is representative of an information management service (IMS) system:

1. The DBMS writes to its log data set; the record written contains old data base (DB) data, new DB data (that is being changed by this transaction), and a record of its intent to commit (finalize) this transaction.

2. The DBMS waits for a DASD I/O operation to report that it has completed, then it updates its data bas data sets, which are diff rent v lumes that ar on a diff r nt DASD subsystems. This writing of the n w DB data overwrites and thus destroys the old DB r cords.

3. The DBMS waits for the second DASD I/O operation to be post d complete, then it writes a commit record to its log data set on the first volume. This commit record 'guarantees' to future IMS recovery processes that the data base data set (DASD volumes) have been updated.

Now consider operation of an asynchronous remote copy system using a multiple subsystem configuration such as illustrated in Figure 1. In this embodiment, the DBMS log data set is on a volume configured to the topmost DASD subsystem pair (DASD 1 and DASD 1') and the data base volumes are on the DASD subsystem pair shown second from the top, (DASD 2 and DASD 2'). In this example, the primary subsystem 1 is lightly loaded, thus it processes its queued work with no delay, while the primary subsystem 2 is heavily loaded and is experiencing some delay in processing its queued work. Queued work includes the forwarding of updated data to its remote copy peer subsystem.

The following sequence would describe the operation in the present example:

1. Write I/O (1) is completed from application to subsystem 1.

2. Write I/O (2) is completed from application to subsystem 2.

3. Primary Subsystem 1 sends data for I/O (1) to secondary subsystem 1, which applies it to its cache memory copy of the DASD volume.

4. Write I/O (3) is completed from application system to subsystem 1.

5. Primary Subsystem 1 sends data for I/O (3) to secondary subsystem 1, which applies it to its cache memory copy of the DASD volume.

6. Primary Subsystem 2 sends data for I/O (2) to secondary subsystem 2, which applies to its cache memory copy of the DASD volume.

If a primary site failure occurs after step 5 and before step 6 there would be corrupted data copied into the system. Since the failure rendered primary subsystems inaccessible, the data from I/O (2) will not be at the secondary site.

The essence of data sequence consistency in a remote copy service is to ensure at such a time as the remote DASD must be used for recovery, that the new data from second operation I/O above will only be seen if the data from the first I/O is also present, and that the data from the third I/O will only be present if the second I/O is also present. Consider the data integrity loss if the third I/O were present but if the second I/O was not. The DBMS log would tell recovery processes that the data base would receive valid data. This would either result in a DBMS failure or business application error.

The sequence of I/O operati ns at the primary subsyst ms is DASD 1, DASD 2, then back to DASD 1; but because of the load on DASD subsystem 2, it is delayed in s nding its I/O (2) such that it has not arriv d by the time I/O (3) from DASD subsystem 1 was

received by secondary DASD subsystem 1'. With no control of transmission subsystem data to copy volum s, subsystem 1 w uld update its copy volume with both 1 and 3. If at that tim a disaster b fell that primary system and op rations w r dir ct d to resume at the secondary site (such contingency being the reason for having a real-time remote DASD copy), the recovering DBMS would find a log record that said that data base records were written (in I/O 2) while the data for I/O 2 would not be on the DASD of secondary subsystem 2.

With such independent links, determination of sequence of writes among independent primary DASD subsystems, communication of that information to the set of subsystems at the secondary location, and control of the sequence of updates among the independent DASD subsystems at the secondary has been a problem heretofore. However, as explained in detail below, a system and method are provided for sequence identification, communication, and update control that permit sequence-consistent remote DASD data copy from multiple independent DASD subsystems at one location to a set of secondary DASD subsystems, each subsystem being independently connected to peer subsystems at the first location.

Figure 2 is a remote dual copy system 20 including a host 11 which provides data to primary subsystems 12. The system 20 includes a first group of DASD subsystems 12 which are located at a primary site and a second group of DASD subsystems 14 which are located at a site that is remote from the primary site. The system 20 includes couplers 16 and 18 which provide for interconnections of the DASD subsystems 12 and 14, respectively.

Referring now to Figure 2, system 20 for achieving update sequence integrity without a centralized communication system is based on the presence of the following configurations.

1. Multiple primary location DASD subsystems each interconnected via one or more communication links to a peer DASD subsystem at the secondary location.
2. A coupling connection 16 of all subsystems at the primary site, and a similar coupling connection 18 of all subsystems at the secondary site. While not shown, one of ordinary skill in the art will readily recognize multiple physical connections may be incorporated for connection redundancy. The term "coupling" is used for convenience; any suitable physical network interconnection means may be used such as a local area n twork (LAN) or the like.
3. Each subsystem has a 'clock' or similar synchronizing signal process that can be synchroniz d with a valu from another subsystem, communicated via the coupling 16.

Thre steps ar utiliz d to achieve copy update s qu nce integrity: (1) determination of sequence of write operations among all DASD subsystems at the primary; (2) communication of that s quence information to the s condary; and (3) use of that information by the sec ndary DASD subsystems to control the sequence of update writes across all secondary DASD. These three steps are described in more detail hereinbelow for a copy system using LAN interconnections at the secondary and primary subsystems.

1. Use of the LAN interconnection among the primary DASDs to distribute a sequencing signal, such as a time clock, to all subsystems is used to associate a sequence/time value with each DASD write of data to be copied to a secondary DASD subsystem (and the sending of that value along with update data to the secondary);
2. Propagation of periodic synchronizing-time-denominated checkpoint signals among the primary DASD subsystems that in turn are communicated by each primary subsystem to its peer-connected secondary subsystem(s); and
3. Use of the LAN interconnection among secondary DASD subsystems to coordinate their DASD update writing of copy data received from primary subsystems.

Referring now to Figure 3, what is shown is a flow chart of the general operation 100 of such a system that is located within the primary subsystem. First, a start copy operation is sent to all the primary subsystems to activate communication between primary subsystems, via step 102. Then individual subsystems provide the appropriate sequence information to all the copy active primary subsystems, via step 104. Thereafter the primary subsystems communicate the sequence information to the peer coupled secondary subsystems, via step 106. Finally, that sequence information is utilized to control secondary subsystem updates via step 108.

These steps are described in detail below:

Start copy operation (step 102)

Remote copy operations for each DASD subsystem must be started by some instruction to that subsystem. The instruction may come via command from a host system (the same system as may be executing applications that write to DASD), or it may be provided via subsystem-local interface such as from a subsystem operator console or similar means.

Description of DASD write sequence at the primary (step 104)

Refer now to Figure 3A and th following discussion. Irrespective of whether the command comes from the host system r from subsystem local interface, the start copy instruction identifies the DASD to be copi d and causes the subsystem to activate com-

munications with secondary subsystem(s) to which it is connected via step 202.

The start copy peration also activates communication from that DASD subsystem to other primary DASD subsystems. The LAN connection addresses of ther primary DASD subsystems may be configured to each subsystem or it may be incorporated in information contained in the start copy instruction to the subsystem. When copy is started at a subsystem, each subsystem sends a subsystem copy active message to all other primary subsystems it has addresses for via step 204. All primary subsystems build and maintain configuration tables such that each subsystem knows the identity of all primary subsystems participating in the copy process via step 206. Each subsystem receiving an copy active message from another subsystem marks that subsystem as a copy active in its configuration tables.

As a part of exchanging copy active messages with the other primary systems, the subsystems synchronize their sequence clock processes and select one subsystem to be a master source for a timing value synchronization signal via step 208. This is a clock synchronization process, not described here since such processes are well known in the art. Note that clock synchronization must be able to maintain clock drift such that maximum drift is substantially less than the time for a subsystem to receive a write command from a system, perform a write to cache, signal end of I/O operation, and for the host to process the I/O completion status and start a new DASD I/O write operation.

As write operations are performed by each subsystem, in a preferred embodiment, the subsystem includes the then current time sequence signal value with other control information that is sent with the DASD data to its connected secondary subsystem. The primary system DASD write I/O operation continues without delay, extended only by the time necessary to construct control information for the data to be sent to the secondary. DASD and control data are buffered in the secondary subsystem on receipt. Update of secondary DASD copy is deferred until released by secondary sequence control (described below).

Communication of sequence information to secondary DASD subsystems (step 106)

Refer now to Figure 3B and the following discussion. The subsystem that is providing the time synchronizing signal source will periodically send a checkpoint message to all primary copy-active subsystems via step 302. This may be included with the time sync signal or sent separat ly as appropriate for the local int rconn ction protocol used. Th checkpoint message includes a s quence time value and a ch ckp int s quence number that is incremented by one for each checkpoint message. Each subsystem on receiving the checkpoint communication will logically insert it in its transmission stream to the secondary subsystem(s) to which it is connected, sending it to secondary subsystem(s) only after all DASD and control information with an earlier sequence time signal value have been sent via step 304.

Use of sequence information by secondary DASD subsystems to control secondary DASD updates (step 108)

Refer now to Figures 3C and 3D and the following discussion. Copy operations in secondary systems are activated by copy active messages from their connected primary subsystems via step 402. (An enabling startup instruction from a system or local console at the secondary may also be required. That aspect of copy operation and control is performed in a conventional manner.) Secondary subsystems build and maintain copy active tables such that each subsystem has data relating to the identity of all secondary subsystems participating in the copy process, via step 404. A synchronizing control master is selected from among the secondary subsystems, using the local interconnect (eg a LAN) similar to the manner in which a primary synchronizing signal source was selected, via step 406. Such distributed control schemes are well known and need not be described here.

Secondary subsystems receive and buffer DASD data and associated control information from their connected primary subsystems. This received data and control information is logically grouped and sequenced by the primary synchronizing signal value. For maximum protection, received data and control information buffering should be in a non-volatile storage medium.

At some point, each subsystem will receive a checkpoint control message via step 408 that signifies that all DASD update data and control with a primary synchronizing signal value equal to or less than the checkpoint time sync value has been sent to that secondary subsystem. A checkpoint message, when received, is sent by the receiving subsystem to the secondary master subsystem, which performs a rendezvous for that checkpoint message from all copy-active secondary subsystems, including itself, via step 410.

When the rendezvous is complete via step 412 for a given checkpoint value, the secondary master subsystem sends a message to all secondary copy-active subsystems to release update data up to the primary sequence time valu of the ch ckpoint. Each subsystem then marks its lf in an int rnal update in progr ss state and applies the updates to its secondary DASD copy via step 414. (Note: The actual process of applying the buff red DASD copy data may re-

quire only adjusting cach directory ntries.)

The update in progress state must be maintained through subsystem resets and pow r off along with buffered data and control info from the primary (and ther copy service state information). It is used as a must-complete type operation control that precludes any host system access to secondary DASD when that state is active. This ensures that an application takeover at the secondary cannot see partial updated DASD data. That is, it can not access a sequence-inconsistent DASD copy. Update for the checkpoint interval must be complete before a user can access the copy DASD subsystems records.

When the updating of DASD copy data has completed, the subsystem sends an update complete for checkpoint message (identifying the specific checkpoint) for that checkpoint to the secondary master, via step 416. When update complete signal for checkpoint messages have been received at the master from all subsystems including the master, the master then sends a reset update in progress state message to all secondary subsystems to allow secondary copy data to again be accessible to attached systems, via step 418.

In a variation to the preferred embodiment, if the coupling means among all primary subsystems can reliably propagate every checkpoint message to all subsystems in substantially less time than the time for an I/O operation cycle then the preceding processes could function without a clock and clock sync. The arrival time of a checkpoint message at each primary subsystem would be precise enough to define update sequence. All updates within a checkpoint would be considered to have occurred at the same time.

The steps of subsystem operation for the three I/Os described previously, using the approach described above is discussed herein below.

1. DASD 1, 2 and 3, and 4 exchange 'copy active' messages. Primary subsystem 3 has become the master source for timing value sync signal. Secondary system 4 has become the secondary master for rendezvous of checkpoint messages.
2. Write I/O (1) is completed from application to subsystem 1 at time 'a'.
3. Write I/O (2) is completed from application to subsystem 2 at time 'b'.
4. Primary Subsystem 1 sends data for I/O (1) to secondary subsystem 1 along with its associated time value 'a'. Secondary subsystem 1 buffers the data but does not apply it to its cache copy for the DASD volume.
5. Subsystem 3 s nds a checkpoint message containing checkpoint sequence number 'n' and time value 'b'.
6. Write I/O (3) is completed from application system to subsystem 1 at time 'c'.
7. Primary Subsystem 1 sends data for I/O (3) to s condary subsystem 1 along with its associated tim valu 'c'. S condary subsystem 1 buffers th data but does not apply it to its cache copy for the DASD volume.
8. Primary subsystem 1 receiv s and processes the checkpoint message sent by subsystem 3 in step 5.
9. Primary subsystem 1 sends checkpoint message to secondary subsystem 1, which forwards it to secondary subsystem 4.
10. Primary Subsystem 2 sends data for I/O (2) to secondary subsystem 2 along with its associated time value 'b'. Secondary subsystem 2 buffers the data but does not apply it to its cache copy for the DASD volume.
11. Primary subsystem 2 receives and processes the checkpoint message sent by subsystem 3 in step 5.
12. Primary subsystem 2 sends checkpoint message to secondary subsystem 2, which forwards it on to secondary subsystem 4.
13. At some point between steps 5 and 13, subsystems 3 and 4 have sent the checkpoint message 'n' to their secondary subsystems. Secondary subsystem 3 has forwarded it to secondary subsystem 4.
14. Secondary subsystem 4 sends a 'release' message to secondary subsystems 1, 2 and 3.
15. Secondary subsystem 3 having no update immediately returns an "update complete" message to secondary subsystem 4.
16. Secondary subsystem 1 enters 'update in progress' state, then applies update (1). It does not apply update (3) since its sync time value 'c' is greater than checkpoint time value 'b'. It then sends an update complete message to secondary subsystem 4.
17. Secondary subsystem 2 enters update state and applies update (2), and sends an update complete message to secondary subsystem 4.
18. Secondary subsystem 4 having received update complete messages from all secondary subsystems, sends a 'reset update in progress state' message to secondary subsystems 1, 2 and 3.

Now if a primary site failure happens at any point in the above sequence the secondary DASD will either show none of the updates, or will show updates (1) and (2). Update (3) will not be 'applied' to secondary subsystem 1's DASD and cache until the next checkpoint sequence.

Accordingly, through the present system a sequence consistent real-time asynchronous copy system is provided that does not require the use of central communications service. In so doing, a system is provided that requir s minimal modification, while utilizing existing capabiliti s within the DASD subsys-

t ms to provide for sequence modifications.

Claims

1. A method of operating an asynchronous remote data copy system (20) including a primary site having a first plurality of subsystems (12) interconnected by a first coupling means (16), and a secondary site remote from the primary site having a second plurality of subsystems (14) interconnected by a second coupling means, each of the second plurality of subsystems being independently coupled to a counterpart one of the first plurality of subsystems, said method comprising the steps of:

   in the first plurality of subsystems;

   sending a checkpoint signal to each of the first plurality of subsystems;

   sending updated data and the checkpoint signal to each of the counterpart coupled second plurality of subsystems; and

   in the second plurality of subsystems;

   receiving the updated data and checkpoint signals; and

   coordinating the writing of the updated data based upon the checkpoint signals.

2. The method of claim 1, further comprising the step of distributing a sequence signal at the primary site, said checkpoint signal having a predetermined relationship with the sequence signal.

3. The method of claim 1 or 2 further comprising the steps of:

   activating each of the first plurality of subsystems to communicate with the other subsystems in said first plurality of subsystems;

   activating each of the first plurality of subsystems to communicate with its counterpart coupled subsystem in the second plurality of subsystems;

   building at least one configuration table in each of the first plurality of subsystems such that each of the first plurality of subsystems can identify all of the other subsystems in said first plurality of subsystems; and

   synchronizing the first plurality of subsystems.

4. The method of any preceding claim in which the coordinating step further comprises the steps of:

   receiving copy active messages in the s cond plurality of subsystems;

   building copy active tables in th s cond plurality of subsystems;

   synchronizing copy op rations of the second plurality of subsystems;

   rec iving th checkpoint messag s in the s cond plurality of subsystems;

   performing a rendezvous for all checkpoint messages in the second plurality of subsystems; and

   applying the updated data at the second plurality of subsystems.

5. The method of any preceding claim, further comprising the steps of:

   causing each of said first plurality of subsystems to asynchronously generate a sequence of updates;

   ordering each of said sequences in accordance with the received sequence signal and checkpoint signal;

   asynchronously communicating sequences of updates from each of the first plurality of subsystems into a buffered portion of a counterpart subsystem; and

   applying the buffered sequences at the second plurality of subsystems as a function of each checkpoint message.

6. A method of operating an asynchronous remote data copy system (20) including a primary site having a first plurality of subsystems (12) interconnected by a first coupling means (16), and a secondary site remote from the primary site having a second plurality of subsystems (14) interconnected by a second coupling means, each of the second plurality of subsystems being independently coupled to a counterpart one of the first plurality of subsystems, said method comprising the steps of:

   (a) at the primary site responsive to initiation of a start copy operation, ascertaining at each subsystem the subset of the plurality of DASD subsystems forming the copyset group, designating one of the plurality of subsystems as a clocking and checkpoint message source, each checkpoint message including a sequence clock value and an increased sequence number;

   (b) at the secondary site, repeating step (a) for counterpart ones of the DASD subsystems;

   (c) at the primary site, periodically generating clocking signals and checkpoint messages by said designated subsystem and broadcasting said signals and messages as they occur to other subsystems forming the copyset group including itself, at each subsystem in the copyset group, and

   (1) asynchronously forming a local sequence of updated records,

   (2) emb dding said signals and messages int the s quenc to f rm a tim discriminated total ordering of updated r cords,

and

(3) remitting at l ast a portion of th sequence to a buffer portion of the counterpart DASD subsyst m at th secondary site; and

(d) at the secondary site, applying a checkpoint message to the designated subsystem operative as a synchronizing source by each subsystem in the copyset group and writing the sequences or portions thereof stored in the buffered portions of said subsystems to DASD only upon a signal from said designated subsystem indicative of its receipt of all send messages.

7. The method according to claim 6, wherein each checkpoint message signifies that all DASD records with a clock signal with a sequence number that is less than the checkpoint message sequence clock value have been transmitted to the counterpart secondary subsystem.

8. An asynchronous remote data copy system (20) including a primary site having a first plurality of subsystems (12) interconnected by a first coupling means (16), and a secondary site remote from the primary site having a second plurality of subsystems (14) interconnected by a second coupling means, each of the second plurality of subsystems being independently coupled to a counterpart one of the first plurality of subsystems,

the first plurality of subsystems including means for sending a checkpoint signal to each of the first plurality of subsystems; and means for sending updated data and the checkpoint signal to each of the counterpart coupled second plurality of subsystems;

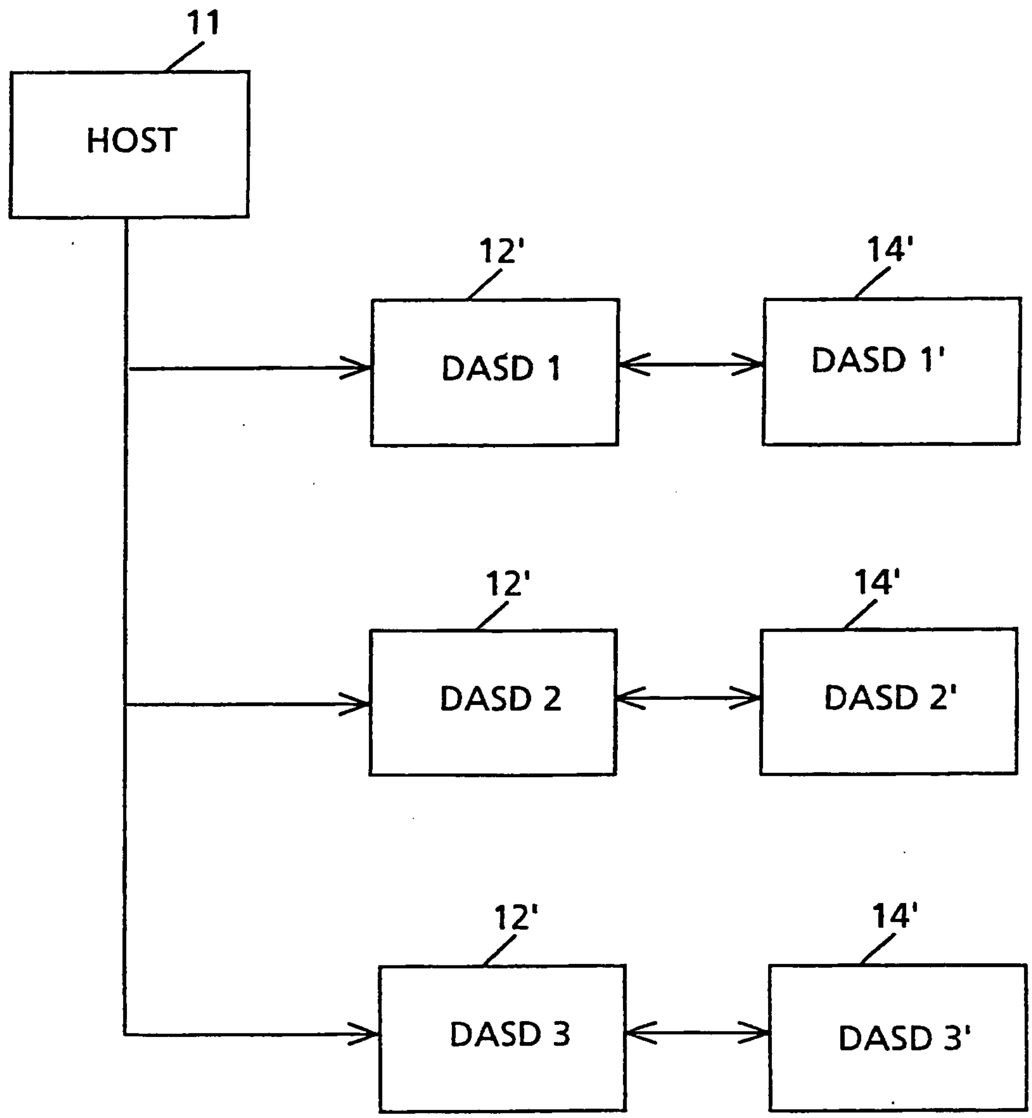
and the second plurality of subsystems including means for receiving the updated data and checkpoint signals; and means for coordinating the writing of the updated data based upon the checkpoint signals.

9. The system of claim 8, further comprising the means for distributing a sequence signal at the primary site, said checkpoint signal having a predetermined relationship with the sequence signal.

10. The system of claim 9 or 10 in which the checkpoint signal sending means comprises:

means for sending a checkpoint message to the first plurality of subsystems; and

means responsive to the checkpoint message sending m ans for inserting the ch ckpoint message into an update data sequence from each of said first plurality of subsystems to its count rpart subsystem.



10
PRIOR ART

Figure 1

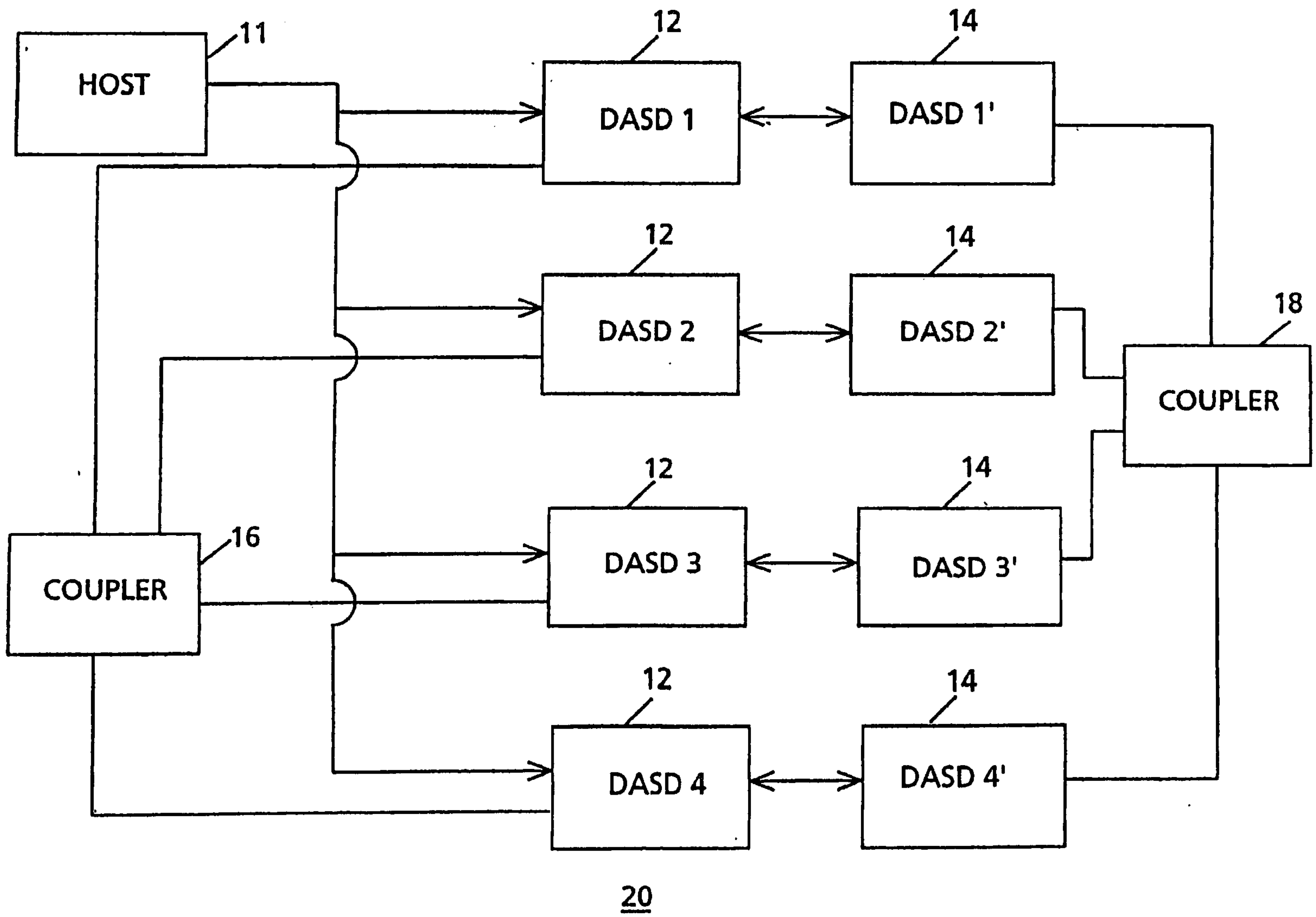

HOST
11
COUPLER
16
DASD 1
DASD 2
DASD 3
DASD 4
12
DASD 1'
DASD 2'
DASD 3'
DASD 4'
14
COUPLER
18
20

Figure 2

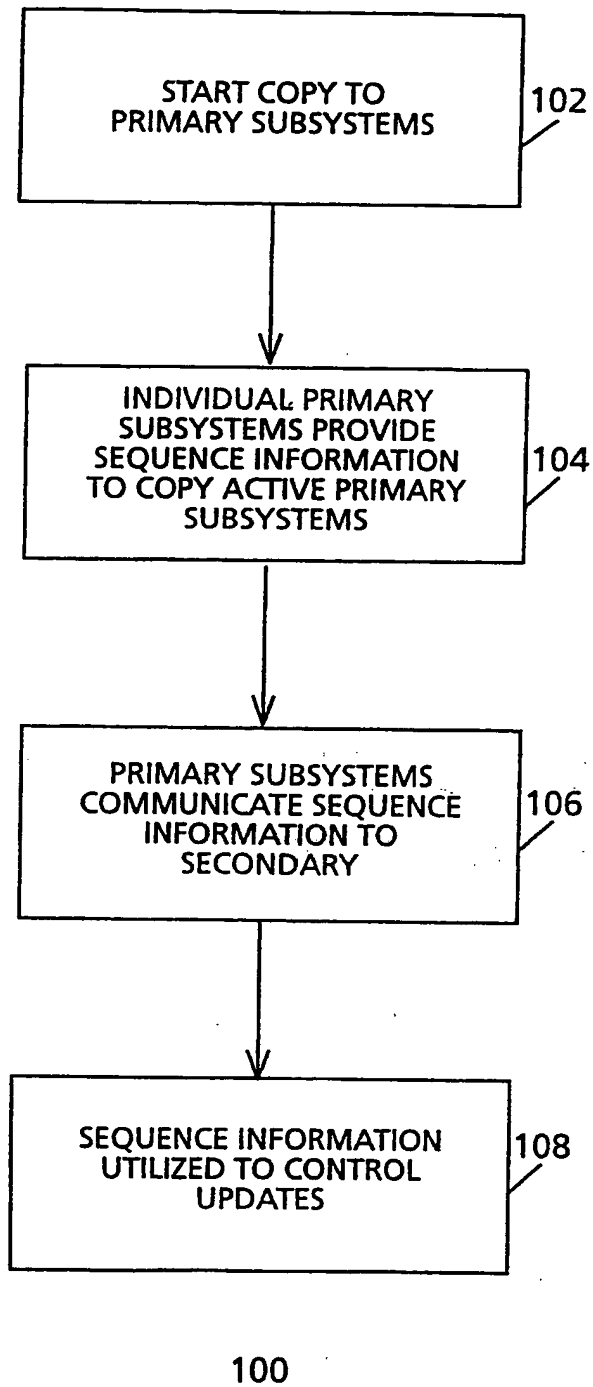

START COPY TO PRIMARY SUBSYSTEMS
102
INDIVIDUAL PRIMARY SUBSYSTEMS PROVIDE SEQUENCE INFORMATION TO COPY ACTIVE PRIMARY SUBSYSTEMS
104
PRIMARY SUBSYSTEMS COMMUNICATE SEQUENCE INFORMATION TO SECONDARY
106
SEQUENCE INFORMATION UTILIZED TO CONTROL UPDATES
108
100

Figure 3

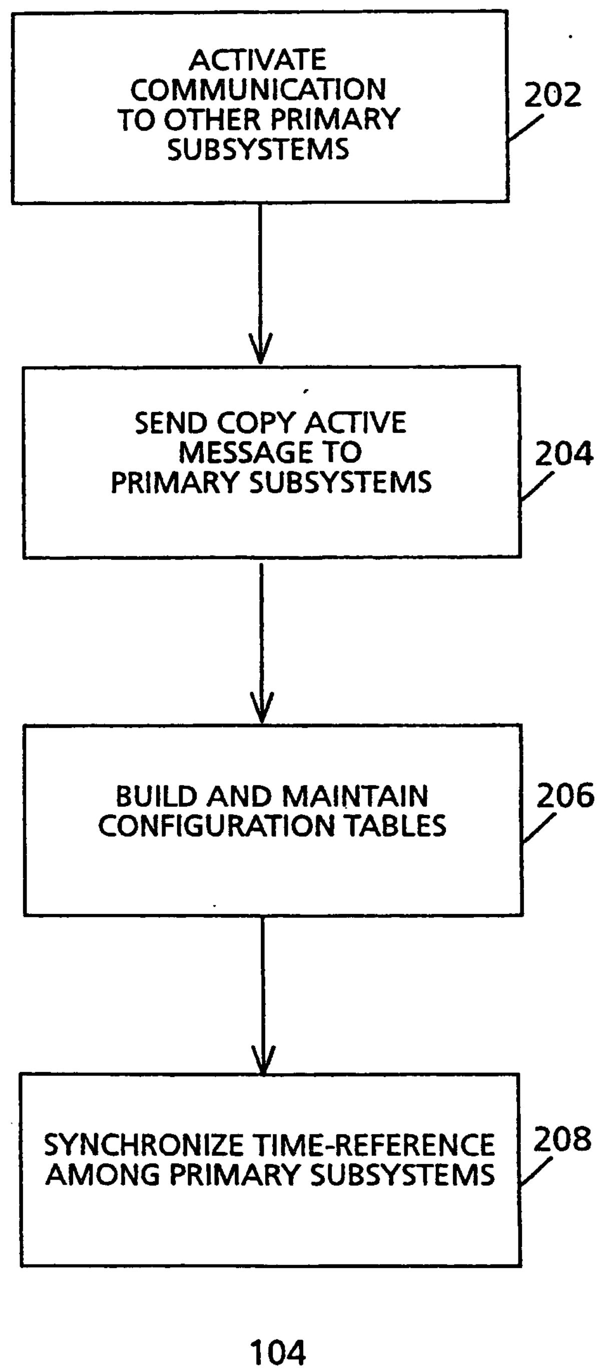

ACTIVATE COMMUNICATION TO OTHER PRIMARY SUBSYSTEMS
202
SEND COPY ACTIVE MESSAGE TO PRIMARY SUBSYSTEMS
204
BUILD AND MAINTAIN CONFIGURATION TABLES
206
SYNCHRONIZE TIME-REFERENCE AMONG PRIMARY SUBSYSTEMS
208
104

Figure 3A

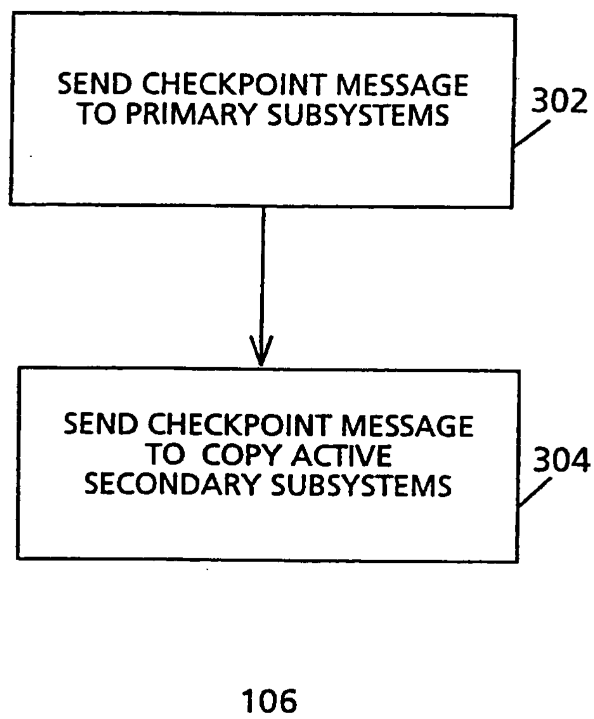

SEND CHECKPOINT MESSAGE
TO PRIMARY SUBSYSTEMS
302
SEND CHECKPOINT MESSAGE
TO COPY ACTIVE
SECONDARY SUBSYSTEMS
304
106

Figure 3B

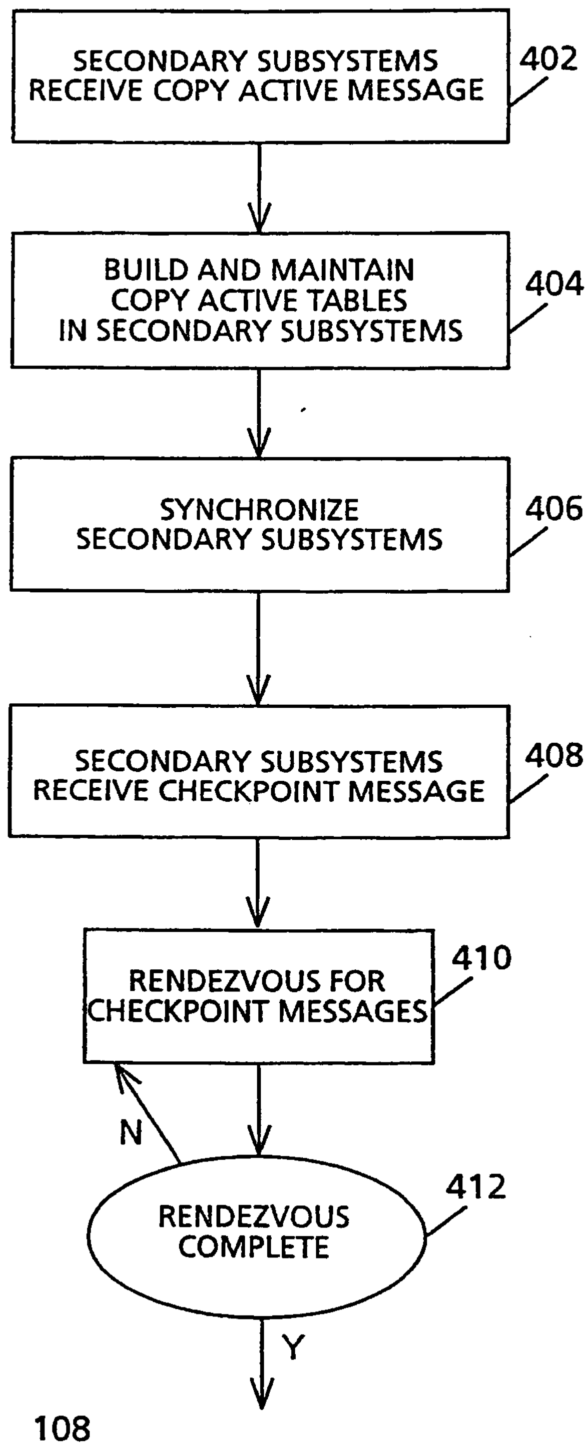

SECONDARY SUBSYSTEMS RECEIVE COPY ACTIVE MESSAGE
402
BUILD AND MAINTAIN COPY ACTIVE TABLES IN SECONDARY SUBSYSTEMS
404
SYNCHRONIZE SECONDARY SUBSYSTEMS
406
SECONDARY SUBSYSTEMS RECEIVE CHECKPOINT MESSAGE
408
RENDEZVOUS FOR CHECKPOINT MESSAGES
410
N
RENDEZVOUS COMPLETE
412
Y
108

Figure 3C

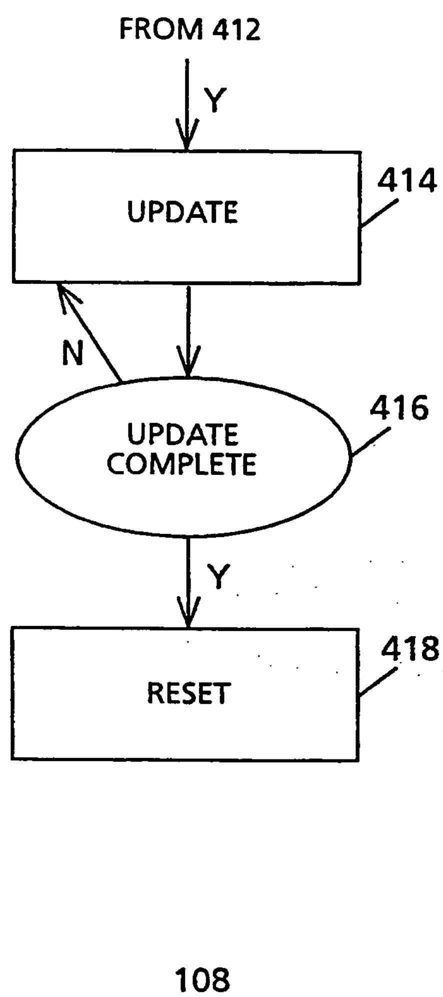


Figure 3D

European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 95 30 1345

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.6)
A,D D	WO-A-91 20034 (STORAGE TECHNOLOGY CORP) 26 December 1991 * abstract * * page 12, line 22 - page 13, line 5 * * page 28, line 31 - page 30, line 16 * * page 45, line 34 - page 47, line 9 * * figure 2 * & US-A-5 155 845 (BEAL ET AL.) ---	1,6,8	G06F11/20 G06F11/14
A	PROCEEDINGS OF THE ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMEN OF DATA, SAN DIEGO, JUNE 2 - 5, 1992, vol. 21, STONEBRAKER M, pages 236-245, XP 000393592 ANUPAM BHIDE ET AL 'AN EFFICIENT SCHEME FOR PROVIDING HIGH AVAILABILITY' * abstract * * page 236, right column, line 7 - line 12 * * page 237, right column, line 32 - line 50 * * page 238, left column, line 46 - right column, line 52 * ---	1,6,8	
			TECHNICAL FIELDS SEARCHED (Int.Cl.6)
			G06F
A	EP-A-0 455 922 (IBM) 13 November 1991 * column 4, line 29 - column 7, line 12 * * column 10, line 28 - line 29 * -----	1,3,6,8	

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
BERLIN	27 June 1995	Masche, C

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons
& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)